

Weekly Report

Junhua Lu

2016 年 7 月 17 日

Done

- Finish Chap. 5 of *DOM Scripting: Web design with JavaScript and the Document Object Model*
- Revised vis papers according to the plan. Much work to be done.
- Miscellaneous: Spent much time on accommodating my dormitory, transport my mass from one dorm to another dorm. I found several bugs while registering for vis conference and also spent much time on it. Assisted Singapore students on their model estimation and evaluation (significance). Introduced our work to a CKC student and instruct her to learn several essential skills. I will assign several tasks for her according to her learning pace. Evaluated a Tianchi assignment.

To do

- Discuss with Guan about how to polish our system design. Proceed to revise according to the new system.
- Apply for a visa in Shanghai.
- Design tasks, and then visit netease, do a user study.
- Go home if time is available.

Papers.

- *Event detection, tracking and visualization in Twitter: A mention-anomaly-based approach*, JOURNAL: Social Network Analysis and Mining. 主要讲的是利用朋友的@关系针对对应twitter 文本挖掘topic. 其可视化是很trivial的在结果方面的可视化, 所以不用多讲, 但是如果整套流程能做成一个系统也是不错的.

文章的重点是这个模型或者说方法, 是通过计算@关系的异常值来判断一件事情是不是一个大新闻. 所谓异常值, 是与一个基本的概率分布假设的差距. 输入就是这些文本, 输出是一个事件列表, 每个事件由(i)一个主词和一系列带权的相关词汇(ii)一个时间段(iii)对人群的重要性. 这个时间段是不固定的, 是一个最大化影响的参数. 事件找出以后, 选择描述词, 这些描述词的选择是通过与主词共现的高低来判断, 由于这是一个时变的数据, 文章采用了一种Erdem等人在2012年提出的correlation coefficient. 这些描述词在下一步存储事件, 去除合并事件中起作用. 整个工作流程如下图 1.

- KDD 2015 *Online Outlier Exploration Over Large Datasets* 这篇文章比较难读, 还没读完, 讲的是一个框架ONION, 用于大型数据集的离群点检测. 这是一种交互式的方法, 是一种不同于平常那种一次

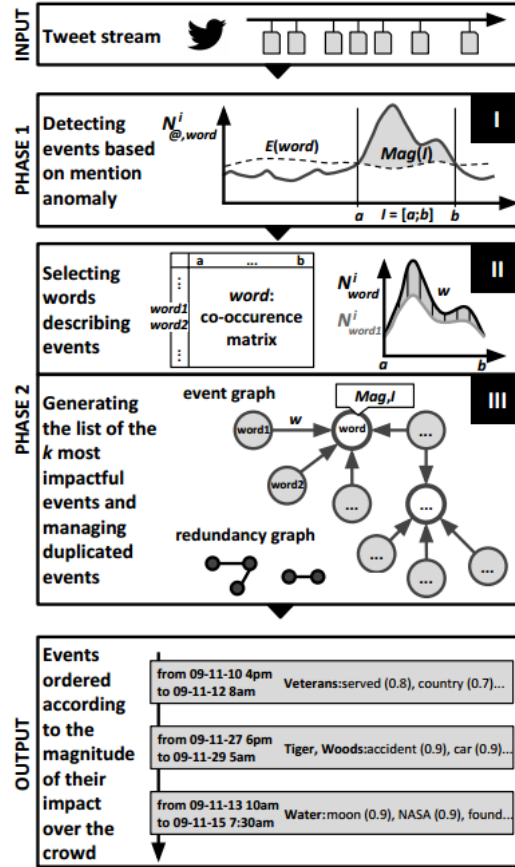


Fig. 1 Overall flow of the proposed method, *MABED*.

图 1: A mention-anomaly-based approach

查询(输入参数)找出来几个这种trial and error的方法. 本文通过对三个不同空间(O空间, 对整套数据以及参数建模; P空间, 参数空间; D空间, 数据空间)的划分以及相互之间的映射关系, 再加上一些排序功能(异常度的排序), 以一个类似于代数的方法呈现出来. 每个空间上都有各自的offline和online的处理过程. 文章理论性比较强, 还没读完, 但是看其实验评估, 不管是在用户交互之后的效率还是实际算法性能都是非常好的.

- 下面读了三篇, 主要是为了读其task-based evaluation的, 顺便粗略看了下其文章的一些可视化内容. 对于evaluation, 大多数文章的套路是这样的: 首先介绍系统, 告诉怎么用; 进行一些简单的上手教学, 哪里不会教哪里, 可能完成基本任务. 在入门以后, 要将我们设计的任务交给他们来做, 常理来说会是几个tasks, 并且会有多个选项(一般为了防止随机, 还会有一个“我不知道”的选项), 记录下做的时间, 答案, 用于做简单的统计分析. 做完系统后, 填个问卷给他打分: 你觉得这个系统怎么样啊, 好用吗, 直观吗, 选起来难吗, 叫你选几个备选你会选这个吗? 然后采访一下. 最后我们手机反馈, 总结. 其实后两者并不是完全并存的, 或者说, 我们当初论文用的哪一套受到了reviewer的质疑, 因此相关部分要缩减, 主要精力仍然是集中在前面的tasks上. Tasks在设计时, 要紧扣两点(1)本身系统的特点, (2)覆盖之前提到的design requirement之类.

值得注意的是, 看了几篇文章都会对系统与其他系统进行比较, 这一点有点难度, 因为我们这个和一般的系统真不好比, 没有一个参照物好说. 这个我们会继续讨论一下. 这三个可视化设计本身都很巧妙, 介绍如下:

CHI16 *Egocentric Analysis of Dynamic Networks with EgoLines* 利用类似地铁图的可视化, 对时变的egonetnetwork进行展现. 每个时间段上整齐排列的地铁图又可以当做邻接矩阵来展现. cluster同色来展示, 一阶邻居, 二阶邻居直观显示; 两个节点如果有联系, 也可以显示的显示一条路径出来, 表现他们是如何关联的.

TVC15 *UnTangle Map: Visual Analysis of Probabilistic Multi-Label Data* 对多标签的item进行可视化, 每个标签有一定的概率. 传统方法是降维等等, 往往是要损失不少信息的. 这个方法用ternary plot不断的堆叠起来, 尽管每次只能三个顶点的label可以展现, 但是其他label可以通过关联性的形式在三角形内部显现. 这样即便是有非常多的标签也能一一展示出来. 并且, 通过关联性, 可以顺藤摸瓜找到最重要的item, 标签等等.

CHI15 *MatrixWave: Visual Comparison of Event Sequence Data* 记得以前有个欧洲皇室家谱图用了类似的设计, 但是本文的设计是借鉴了这个方法, 用于时序数据. 原本sankey复杂的交错可能需要用一些layout算法来减少, 现在用了这个方法直接不需要这样的忧虑了. 当然如果时间段特别长的话, 就又是另外一回事了.